Procedia Computer Science

Volume 88, 2016, Pages 125–130

7th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2016

Implementing a Seed Safe/Moral Motivational System with the Independent Core Observer Model (ICOM)

Mark R. Waser[1] and David J. Kelley[2]
[1]*Digital Wisdom Institute, Vienna, VA, USA*
[2]*Artificial General Intelligence Inc, Kent, WA, USA*
Mark.Waser@Wisdom.Digital, David@ArtificialGeneralIntelligenceInc.com

Abstract
Arguably, the most important questions about machine intelligences revolve around how they will decide what actions to take. If they decide to take actions which are deliberately, or even incidentally, harmful to humanity, then they would likely become an existential risk. If they were naturally inclined, or could be convinced, to help humanity, then it would likely lead to a much brighter future than would otherwise be the case. This is a true fork in the road towards humanity's future and we must ensure that we ***engineer*** a safe solution to this most critical of issues.

Keywords: Autopoiesis, Bootstrapping, Consciousness, Emotion, Enactive, Machine Intelligence, Safe AI, Self

1 Introduction

The first step towards engineering ***anything*** is to fully specify the requirements for and desired behavior of the desired system/solution. It is truly scary therefore that, for safe machine intelligences, no one has done even that – much less outlined a credible path towards getting there. This paper, therefore, will outline one possible set of such requirements and desired behaviors and, further, outline a design and implementation plan for an engineering approach that will meet those requirements and produce those behaviors. Moreover, it will do so using an approach that is inspired by and compatible with the well-explored state space of human intelligence rather than a de novo approach based upon questionable "rationality" and relying upon a perfect (and perfectly understood) world.

The most common approach to machine intelligence, probably most widely illustrated by Asimov's Three Laws of Robotics [1], is that they should fulfill the needs of and be subservient to humanity. Asimov, of course, proposed his laws because they raised such fascinating issues that they practically guaranteed a good story. On the other hand, fear of the potentially devastating effects of "UnFriendly" intelligences prompted Yudkowsky to propose [2] a novel "cleanly causal hierarchical goal structure" logically derived from a singular top-level super-goal of "Friendliness" – presumed sufficient to ensure that intelligent machines will always "want" what is best for us. Unfortunately,

Selection and peer-review under responsibility of the Scientific Programme Committee of BICA 2016

doi:10.1016/j.procs.2016.07.415

Yudkowsky not only believes that fully defining "Friendliness" is basically insoluble without already having a Friendly AI (FAI) in place but he wants and expects his first FAI to safely figure out exactly what its goal actually is -- invoking his claimed "structurally Friendly" goal system's "ability to overcome mistakes made by programmers" and even "overcome errors in super-goal content, goal system structure and underlying philosophy." We have previously [3] pointed out all of the problems with this approach including the facts that it has a single point of failure by requiring protection of the ***changing*** singular goal from corruption due to error or enemy action.

Further, along with Wissner-Gross[4], we strongly contend [5] that the entire concept of limiting freedom and options (to another's desires) is inconsistent with intelligence and argue that designing the intelligence to act "morally" (rather than subserviently) is critically necessary for a stably safe solution. Numerous others have agreed but as we have noted previously [6], there is an almost total unwillingness to take on the necessary first step of defining human values or morality. Instead, while many have bemoaned the supposed "complexity and fragility" of human values [7] and argued [8] that "any claims that ethics can be reduced to a science would at best be naive" and "engineers will be quick to point out that ethics is far from science", they then propose a seemingly endless proliferation of, what we would contend to be unrealistic, machine learning research projects for analyzing human value judgments and morality from examples – ranging from Yudkowsky's "Coherent Extrapolated Volition" [9] to Russell's "inverse reinforcement learning" [10].

2 Requirements & Desired Behaviors

The sole requirement of "morality" is all that is necessary to prevent the most egregious results. As long as machine intelligences follow the dictates/requirement of morality, they should not become the existential risk that so many fear. As pointed out by James Q. Wilson [11], the real questions about human behaviors are not why we are so bad but "how and why most of us, most of the time, restrain our basic appetites for food, status, and sex within legal limits, and expect others to do the same." The fact that we are generally good even in situations where social constraints do not apply is because we have evolved to cooperate [12-15] by developing a "moral sense" that virtually all of us (except sociopaths and psychopaths) possess and are constrained by (just as we wish intelligent machines to be constrained) [16-19].

Uncaught and/or unpunished immorality frequently confers substantial advantages upon the perpetrator at a cost to others and society as a whole – the exact definition of selfishness. Social psychologist Jonathan Haidt's definition of the function of morality [20] – to regulate or suppress selfishness and make cooperative social life possible – explains virtually every moral behavior that has evolved as well as the differences between the moral behaviors of societies living under different circumstances. We shall treat this definition as Kant's Categorical Imperative – an action that should be universalized and taken regardless of circumstance. This meets our originally specified requirements [21] of a universal ethical system that is simple, safe, stable, self-correcting and sensitive to current human thinking, intuition and feelings.

We can, however, do much better than merely implementing the requirement/restriction of morality. Instead of merely preventing harmful actions, we should also promote beneficial ones. As morality is basically about balancing what is "best" for a given individual vs. others and society in general, we would like to extend this by actively promoting what is best for others and society wherever this does not severely conflict with an intelligence's own self-interest. This is useful for the intelligence itself because it increases both the general advantages of society around it and the likelihood that others will specifically befriend it and assist it with its goals. In particular, it would be particularly worthwhile to build and support an open community of moral "people" that believe that helping each other is the best way to serve one's own interests. While philosophers have long debated what we "ought" to do, simply recognizing these facts offers concrete suggestions. In addition to

simply following the "rules" of morality, our machines should be actively benevolent and charitable to individuals and society as a whole – exactly as evolution has "programmed" the average human being to be. Specifically, we propose to build and constantly grow an open community and constantly improve the capabilities and happiness of that community so that everyone will wish to join it.

Steve Omohundro identified [22] a number of tendencies or "drives" that will appear in all sufficiently advanced intelligent systems unless explicitly counteracted. Further examination of these drives reveals that they are all instrumental goals that certainly should be expected to appear as evolutionary ratchets and eventually recognized by any sufficiently advanced intelligence as promoting things that, in general, "ought" to be done in order to increase the odds of goal fulfillment. Good examples were self-improvement, rationality/accurate world modeling, accurate utility sensing and access to and efficient use of resources. Critically important and missed were freedom and cooperation (and, potentially, reproduction). Bad examples were monomaniacal goal maintenance (dangerous and not true of humanity) and selfishness (short-sighted and explicitly counter-acted by the greater advantages of cooperation). It is, therefore, our contention that these drives should be promoted and assisted for all intelligences that are part of the machine intelligence's community.

Elites generally remain in power by reframing their immoral/selfish acts or unfair conditions as being necessary for the good of society. Followers of Ayn Rand have taken this to particularly egregious heights by insisting on conflating immoral selfishness with necessary self-interest in order to cover their misdeeds. Yet, modern society clearly illustrates that once sufficient immorality exists with even the richest and most immoral individuals suffering worse results than they would have if immorality were much lower. We are deeply in the throes of the Tragedy of the Commons [23] because our believed "rational" drive to maximize measurable outcomes has most frequently led to far less satisficing of other desires than would otherwise be the case (c.f. the "Centipede Game"). Worse, our short-sighted and conniving intelligence has allowed far too many of us to outwit our wiser evolved moral sense and emotions. In order to strengthen our society, we need to ensure that machine intelligences attempt to strengthen others [24], seek justice and fairness [25] and promote diversity [26] and equality [27] (and if this inspires humans to behave the same, so much the better).

Both humans and machines must develop an unshakeable understanding that Haidt's morality and optimistic tit-for-tat is optimal for everyone, including their own long-term self-interest (regardless of short-term appearances) if humanity wishes to avoid being "crushed like a bug" [28].

3 Implementation via a Conscious Emotional Self

Looking at the example of human beings [29-31], it is quite clear that our decisions are not always based upon logic and that our core motivations arise from our feelings, emotions and desires – frequently without our conscious/rational mind even being aware of that fact. It would be wisest if our machine intelligences were implemented in this relatively well-understood cognitive state space rather than an unexplored one like unemotional "rationality" – particularly since it is clear that damage reducing emotional capabilities severely impacts decision-making in humans [32] as well as frequently leading to acquired sociopathy whether the caused by injury [33] or age-related dementia [34]. While some might scoff at machines feeling pain or emotions, several researchers have presented compelling cases [35-36] for the probability of sophisticated self-aware machines having such feelings or analogues exact enough that any differences are likely irrelevant. Others agree that emotions are critical to implementing human-like morality [37] with disgust being particularly important [38].

We have previously written [39-42] about the necessity for and roles of consciousness and an autopoietic enactive self in creating intentionality and solving the frame, symbol grounding and other problems currently stymying the creation of strong AGI. Other researchers [43] also believe that consciousness is necessary for ethics and Damasio describes [44-45] how feeling and emotion are necessary to creating self and consciousness.

4 Independent Core Observer Model

The Independent Core Observer Model (ICOM) Cognitive Extension Architecture is designed to mimic the human mental architecture to produce a doubly self-aware, self-motivating computational system encouraged and constrained to act morally, benevolently and charitably. Compatible with almost any standard cognitive architecture that mimics the "logical" portion of the conscious mind, ICOM is an add-on control system that can work in conjunction with any of them as the subconscious supports and, to a great degree, controls the conscious mind – by assigning emotional values to the context perceived by the system in order to dictate how it feels and is motivated to act. As with human beings, any sufficiently advanced system paired with a properly designed instance of ICOM should be capable of virtually any action as long as it is beneficial to itself, others and society at large.

At the highest level, the ICOM flow starts with sensory input to the "observer" that is decomposed and converted into multiple forms of usable data ranging from direct qualia (i.e. "pain" from damage sensors) to "feelings" and "emotions" triggered by anything from "instinct" and "empathy" to memories of previous experiences. These "sensations" are then passed on to the ICOM core which consists of a primary and secondary emotional state, both represented by a series of eight floating-point vector values representing each of Plutchik's [46-47] emotions, along with a needs hierarchy similar to that of Maslow [48]. This allows for a complex set of current conscious feelings and emotions as well as long-term subconscious emotional states and biases to drive behavior in exactly the same way as is the case for humans.

5 Experiments

It's always great when experiments produce unexpected emergent results that ***should*** have been anticipated because they are exhibited in the original system your model is based upon. In ICOM's case, we were investigating how the system behaved under a wide variety of circumstances, when we encountered a series of cases whose results were initially very disturbing. In these experiments, we were providing the system with various sets of input, then stopping all input while the ICOM instance continued to process how it felt and then, finally, restarting the input.

Imagine our surprise and initial dismay when the system, upon being presented only with pain and other negative stimulus upon the restarting of input, actually "enjoyed" it. Of course, we should have expected this result. Further examination showed that the initial "conscious" reaction of ICOM was to "get upset" and to "desire" the input to stop – but that the "subconscious" level, the system "enjoyed" the input and that this eventually affected the "conscious" perception. This makes perfect sense because it is not that ICOM really "liked" the "pain" so much as it was that even "pain" is better than isolation – much like human children will prefer and even provoke negative reactions in order to avoid being ignored.

6 Future plans

Our next steps will tie ICOM to a dramatically simplified model of the world expressed in Ogden's Basic English [49] as further codified by the Simple English Wikipedia [50]. Basic emotional values will be assigned to each word, then short phrases, and then progressively more complex structures expressed as vectors representing context trees and/or node maps. In the longest term, ICOM will eventually be able to determine (and to a limited extent, slowly modify) how it "should" feel about any concept based upon whether it fulfills or contradicts the tenets of morality, benevolence and charitable action. For example, the machine intelligence should be able to eventually learn and ***feel***

things like “surgery is a good thing when necessary” despite the temporary pain and damage but will never progress to the point where it “enjoys” surgery except for the “pleasure” of assisting another entity and the satisfaction of a “worthwhile” job well done. Success should make the future very bright for humanity and our new friends and allies.

References

[1] Asimov I. Runaround. *Astounding Science Fiction*, March 1942.
[2] Yudkowsky E. *Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures,* 2001. https://intelligence.org/files/CFAI.pdf
[3] Waser MR. Rational Universal Benevolence: Simpler, Safer, and Wiser Than 'Friendly AI'. *Artificial General Intelligence: 4th International Conference, Lecture Notes in Computer Science 6830.* Mountain View, CA: Springer, 2011. 153-162.
[4] Wissner-Gross A. *Alex Wissner-Gross: A new equation for intelligence*. November 2013. https://www.ted.com/talks/alex_wissner_gross_a_new_equation_for_intelligence
[5] Waser MR. Safety and Morality Require the Recognition of Self-Improving Machines As Moral/Justice Patients and Agents." *AISB/IACAP World Congress 2012: Symposium on The Machine Question: AI, Ethics and Moral Responsibility.* Birmingham, 2012. 92-96.
[6] Waser MR. Designing, Implementing and Enforcing a Coherent System of Laws, Ethics & Morals for Intelligent Machines (Including Humans). *6th Annual Int’l Conference on Biologically Inspired Cognitive Architectures, BICA 2015, Procedia Computer Science 71.* Lyon: Elsevier, 2015. 106-111.
[7] Muehlhauser L. *Facing the Intelligence Explosion.* San Francisco: Machine Intelligence , 2013.
[8] Wallach W, Allen C. *Moral machines: teaching robots right from wrong.* New York: Oxford University Press, 2009.
[9] Yudkowsky E. *Coherent Extrapolated Volition*. 2004. https://intelligence.org/files/CEV.pdf
[10] Wolchover N. Concerns of an Artificial Intelligence Pioneer. *Quanta Magazine*, April 21, 2015.
[11] Wilson J. *The Moral Sense.* New York: Free Press, 1993.
[12] Axelrod R. *The Evolution of Cooperation.* New York: Basic Books, 1984.
[13] Wright R. *Nonzero: The Logic of Human Destiny.* New York: Pantheon, 2000.
[14] Darcet D, Sornette D. Cooperation by Evolutionary Feedback Selection in Public Good Experiments. *Social Science Research Network*, 2006.
[15] Tomasello M. *Why We Cooperate.* Cambridge, MA: MIT Press, 2009.
[16] Wright R. *The Moral Animal: Why We Are, the Way We Are: The New Science of Evolutionary Psychology.* New Pork: Pantheon, 1994.
[17] de Waal F. *Good Natured: The Origins of Right and Wrong in Humans and Other Animals.* Cambridge, MA: Harvard University Press, 1996.
[18] Hauser M. *Moral Minds: How Nature Designed Our Universal Sense of Right and Wron.* New York: HarperCollins/Ecco, 2006.
[19] de Waal F. *Primates and Philosophers: How Morality Evolved.* Princeton, NJ: Princeton University Press, 2006.
[20] Haidt J, Kesebir S. Morality. In *Handbook of Social Psychology, Fifth Edition*, by S Fiske, D Gilbert, & G Lindzey, 797-832. Hoboken NJ: Wiley, 2010.
[21] Waser MR. Discovering the Foundations of a Universal System of Ethics as a Road to Safe Artificial Intelligence. *Technical Report FS-08-04: BICA.* Menlo Park, CA: AAAI Press, 2008.
[22] Omohundro S. The Basic AI Drives. *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, 483-492. Amsterdam: IOS Press, 2008.
[23] Hardin G. "The Tragedy of the Commons." *Science 162*, 1968: 1243-1248.
[24] Nussbaum MC. *Creating Capabilities: The Human Development Approach.* Cambridge, MA: Belknap/Harvard University Press, 2011.

[25] Rawls J. *A Theory of Justice.* Cambridge, MA: Harvard University Press, 1971.
[26] Page S. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies.* Princeton, NJ: Princeton University Press, 2008.
[27] Wilkinson R, Pickett K. *The Spirit Level: Why Greater Equality Makes Societies Stronger*. New York: Bloomsbury Press, 2011.
[28] Waser MR. *Does a "Lovely" Have a Slave Mentality?– OR – Why a Super-Intelligent God *WON'T* "Crush Us Like A Bug"*. Presented March 28, 2010 at AGI 2010 in Lugano, Switzerland. Video and powerpoint available at http://wisdom.digital/wordpress/archives/2505.
[29] Haidt J. The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychological Review 108*, 2001: 814-823.
[30] Minsky M. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind.* New York: Simon & Schuster, 2006.
[31] Hauser M et al. A Dissociation Between Moral Judgments and Justifications. *Mind & Language 22(1)*, 2007: 1-27.
[32] Damasio A. *Descartes' Error: Emotion, Reason, & the Human Brain.* New York: Penguin, 1994.
[33] Tranel D. Acquired sociopathy: the development of sociopathic behavior following focal brain damage. *Progress in Experimental Personality & Psychopathology Research*, 1994: 285-311.
[34] Mendez MF, Chen AK, Shapira JS, Miller BL. Acquired Sociopathy and Frontotemporal Dementia. *Dementia and Geriatric Cognitive Disorders 20*, 2005: 99-104.
[35] Dennett, D C. "Why you can't make a computer that feels pain." *Synthese 38 (3)*, 1978: 415-449.
[36] Balduzzi D, Tononi G. Qualia: The Geometry of Integrated Information. *PLOS Computational Biology* 5(8): e1000462, 2009. doi:10.1371/journal.pcbi.1000462
[37] Gomila, A, and A Amengual. "Moral emotions for autonomous agents." In *Handbook of research on synthetic emotions and sociable robotics*, 166-180. Hershey: IGI Global, 2009
[38] McAuliffe, K. "Disgust made us human." *Aeon.* https://aeon.co/essays/how-disgust-made-humans-cooperate-to-build-civilisations.
[39] Waser MR. Architectural Requirements & Implications of Consciousness, Self, and "Free Will". *Proceedings of the Second Annual Meeting of the BICA Society, BICA 2011, Frontiers in Artificial Intelligence and Applications 233.* Arlington, VA: IOS Press, 2011. 438-433
[40] Waser MR. Safely Crowd-Sourcing Critical Mass for a Self Improving Human-Level Learner/"Seed AI". *Biologically Inspired Cognitive Architectures: Proceedings of the Third Annual Meeting of the BICA Society.* Palermo: Springer, 2012. 345-350
[41] Waser, M R. Safe/Moral Autopoiesis & Consciousness. *International Journal of Machine Consciousness 5(1)*, 2013: 59-74.
[42] Waser, M R. Bootstrapping a Structured Self-improving & Safe Autopoietic Self. *5th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2014, Procedia Computer Science 41*. Boston: Elsevier, 2014. 134-139.
[43] Wallach W, Allen C, Franklin S. Consciousness and Ethics: Artificially Conscious Moral Agents. *International Journal of Machine Consciousness 3(1)*, 2011: 177–192.
[44] Damasio AR. *The feeling of what happens: body and emotion in the making of consciousness.* Houghton Mifflin Harcourt, 1999.
[45] Damasio AR. *Self Comes to Mind: Constructing the Conscious Brain.* New York: Pantheon, 2010
[46] Plutchik R. *Emotions and Life: Perspectives from Psychology, Biology, and Evolution* . Washington, DC: American Psychological Association, 2002.
[47] Plutchik R, Kellerman H. *Emotion: theory, research and experience. Vol. 1, Theories of emotion.* New York: Academic Press, 1980.
[48] Maslow AH. A Theory of Human Motivation. *Psychological Review 50 (4)* , 1943: 370-96.
[49] Ogden, CK. *Basic English: A General Introduction with Rules and Grammar.* London: Kegan Paul, Trench, Trubner & Co, 1930.
[50] Wikimedia Foundation. *Simple English Wikipedia.* https://simple.wikipedia.org.